

RATE-DISTORTION-CONSTRAINED STATISTICAL MOTION ESTIMATION FOR VIDEO CODING

Wilson C. Chung, Faouzi Kossentini, and Mark J. T. Smith

Digital Signal Processing Laboratory
 School of Electrical & Computer Engineering
 Georgia Institute of Technology
 Atlanta, Georgia 30332-0250, USA

ABSTRACT

A rate-distortion-constrained statistical motion estimation algorithm is presented here that leads to improvements in subband video coding. The main advantages of the algorithm is that it requires a relatively small number of computations, produces a much smoother motion field, and employs a more effective measure of performance than the conventional mean absolute difference or mean squared error. The proposed algorithm circumvents problems in the motion compensation loop such as illumination variations, noise, and occlusions, by providing a mechanism for alternating between intra-frame and residual coding. Experimental results demonstrate that the corresponding video coder outperforms the H.263 in terms of motion vector search complexity and overall bit rate at the same reproduction quality.

1. INTRODUCTION

In conventional video coding systems, block matching algorithms (BMAs) are often used for motion estimation to remove temporal redundancies [1, 2, 3]. Such algorithms form the foundation for many video coders and are part of the H.261, H.263, and MPEG standards [4, 5, 6, 7], mainly because they are relatively simple in concept and design, but also because they tend to work reasonably well.

A disadvantage of BMAs, in general, is that their performance is sensitive to illumination changes, noise, occlusion, and reconstruction quality of previously coded frames. Motion vector estimates often do not correspond to physical motion in the video scene. Even where motion does not exist, BMAs produce an estimate. This can lead to a rough motion field, where many motion vectors carry little useful information, yet are very difficult to encode. Moreover, since a mean squared error or mean absolute difference distortion measure is usually used as the matching criterion, the motion vector estimate does not necessarily lead to the best rate-distortion performance [8, 9, 10]. To address some of these problems, the MPEG-2 standard, for example, provides a mechanism for alternating between intra-frame and inter-frame coders.

In this paper, we introduce a rate-distortion constrained statistical motion estimation algorithm that not only requires a level of complexity that is comparable to that of

This work was supported in part by the Joint Services Electronics Program (JSEP).

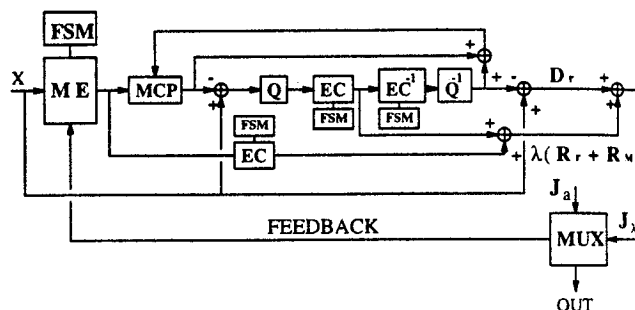


Figure 1: Block diagram of rate-distortion-constrained statistical motion estimation.

the fastest BMAs but also solves most of the above problems, thereby leading to a more consistent motion field. The algorithm is used for motion estimation on the subband level [11, 12], where high order entropy-constrained residual scalar quantization is employed for coding both the original and residual subbands. The proposed algorithm exploits the natural motion field smoothness that tends to exist spatially, temporally, and across subbands. It selects motion vectors based on the current behavior of the motion field and also based on the performance of the residual coder, which is also the ultimate objective performance measure of the video coder. Although the proposed algorithm is presented in the context of subband video coding, its underlying principles can also be applied in other contexts.

2. THE PROPOSED MOTION ESTIMATION ALGORITHM

The proposed motion estimation algorithm is illustrated in Fig. 1. Each frame of the video sequence is decomposed into M subbands using a uniform subband decomposition structure. The algorithm is applied to each subband independently, using information from previously coded subbands (see Fig. 2). First, sufficiently large block and search region sizes are chosen for each subband. All of the possible motion vectors in the search region are then divided into clusters or rectangular regions. This is illustrated in Fig. 3, where each black dot denotes a motion vector location. The search region shown in the figure corresponds to ± 4 displacements for each of the two coordinates. During the

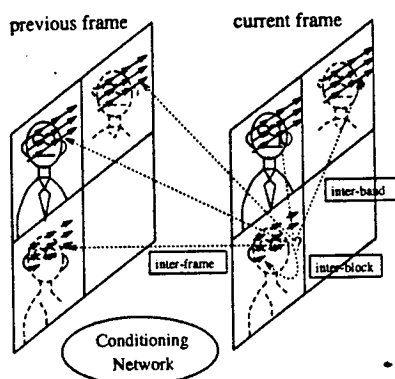


Figure 2: Inter-frame, inter-subband, and intra-subband dependencies for motion vectors.

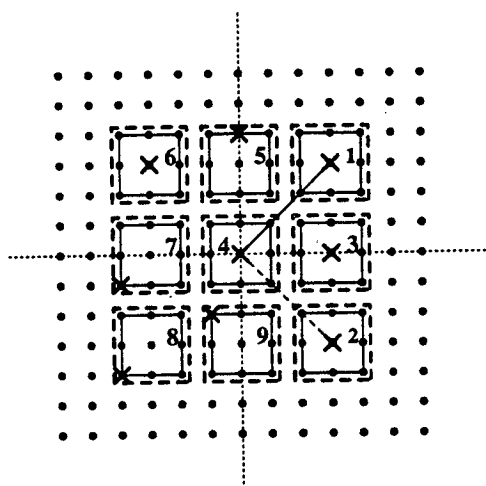


Figure 3: An example of first layer and second layer passes in intra-band motion estimation.

design process, conditional probabilities are generated for each of the rectangular regions, and these probabilities are grouped into tables, each corresponding to a conditioning state. The states are derived based on previously coded motion vector in a subband-spatial region of support. In other words, a high order statistical model that is driven by a finite-state machine (FSM) is built that exploits statistical dependencies in the motion field between motion vectors within the current subband as well as between subbands in both the same frame and previous frames. Complexity reduction techniques described in [13] allow us to use a sufficiently large conditioning subband spatio-temporal region of support, yet produce only a small number of conditioning states. Since the motion field is usually quasi-stationary, adaptation is used during the encoding procedure, but the conditioning network is kept fixed for a larger number of frames.

Given a conditioning state, the algorithm, illustrated in Fig. 3, performs two passes. In the first layer pass, the motion vector (solid line terminated by \times) with largest probability p_i contained in the rectangular region with the largest conditional probability p_j (i.e. region 1 in Fig. 3)

is selected first as the candidate motion vector. A high order entropy-constrained residual coder [13] is then applied to the difference between the original block and the motion-compensated prediction block, producing a rate R_r and a distortion D_r , as shown in Fig. 1. Next, we compute the Lagrangian $J_\lambda = D_r + \lambda(R_m + R_r)$, where R_m is the motion vector bit rate, set here for simplicity to the sum of conditional self-information components ¹ $R_m = -\log_2(p_i) - \log_2(p_j)$. Let J_a be the current running average Lagrangian and T_1 be a threshold² that determines the tradeoffs between complexity and rate-distortion performance. If $J_\lambda \leq T_1(J_a)$, then the selected motion vector is accepted and encoding is terminated for that block by sending motion and residual encoded bits to the channel. At this point, practically no computations have been performed for the estimation procedure. All multiplies/adds performed would have been needed for encoding subsequently. If $J_\lambda > T_1(J_a)$, then the selected motion vector is rejected and a signal is fed back to the motion estimator, where the most probable motion vector located in the region with the second largest conditional probability p_j is selected as an alternative candidate. This is indicated by the dashed line terminated by \times in region 2 of Fig. 3. This procedure is repeated until either the above condition is met, when encoding is aborted, or all regions are exhausted. In cases where little or no motion exists in the video scene, encoding is aborted in the early stages of the first layer pass. However, in cases where the video signal undergoes sudden changes (e.g., zoom, occlusion, illumination), accurate motion vectors cannot be predicted based on the probabilities in the model because no *a priori* information about sudden motion variations is available. As a result, an inaccurate motion vector predicted by the statistical model will generally lead to an increase in the Lagrangian value J_λ . In such cases, a second layer pass is employed.

In the second layer pass, the lowest Lagrangian J_λ^* is compared to J_a . If $J_\lambda^* > T_2(J_a)$, where T_2 is a threshold, whose best value is found experimentally to be between 2.0 and 3.0, then the algorithm exits. Otherwise, the region that led to the lowest Lagrangian is again considered, where other less probable motion vectors belonging to the same region are chosen as candidates. The algorithm proceeds by applying the same procedure as in the first layer pass. In other words, for the next most probable motion vector, the new motion-compensated prediction block is computed, and the same entropy-constrained residual scalar coder is applied to the corresponding residual block. The same procedure is repeated until the proper condition is satisfied, or all regions are exhausted. Finally, in the case where the algorithm exits the two passes without yielding any "good" motion vector candidate, the lowest Lagrangian produced during both passes is compared to that of the intra-frame coder, and the coder leading to the lower value is used. Details of the rate-distortion-based mechanism, by which a particular coder is chosen, as well as a complete description of the residual coder can be found in [12].

¹Note that by storing $-\log_2(p)$ instead of a probability p , no \log_2 operations need to be performed.

²The best value of T_1 is determined experimentally, and is usually between 1.0 and 1.5.

3. ADVANTAGES OF THE PROPOSED ALGORITHM

At first glance, the proposed motion estimation algorithm seems quite complicated. However, experimental results show that, depending on the bit rates of operation and the contents of the video scene, our algorithm stops after the first stage of the first layer pass, which requires practically no computations, more than 90% of the time. Moreover, the algorithm completes both passes in only approximately 2% of the cases. Besides its computational advantage, the proposed algorithm has several other features worthy of mention. First, it efficiently and effectively exploits dependencies between motion vectors by using a two-layer, region-based and vector-based, statistical model. Second, it improves the consistency of the spatio-temporal smoothness of the motion field and reduces sensitivity of the estimation by favoring the most probable candidates. To illustrate this, Figures 4 (a) and (b) show the comparison between the motion fields resulted from our algorithm and the full-search BMA. The increased smoothness observed in Fig. 4 (b) indicates a lower entropy. Not only it is easier to encode the motion vectors in Fig. 4 (b), but the subjective quality of the reconstructed video frames is also better. Another advantage of our algorithm is that it directly embeds the residual coder into the estimation loop, thereby potentially leading to overall better rate-distortion performance. Finally, note that the number of computations required by the proposed motion estimation algorithm is variable, and depends on the content of the video scene. In variable length video coders (such as MPEG), this can be incorporated into the buffering schemes already being used in bit rate control.

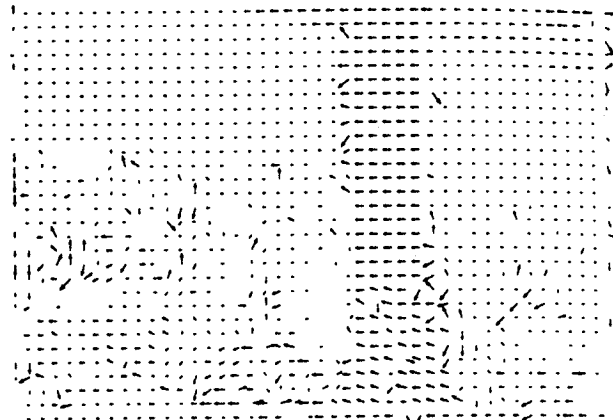
4. EXPERIMENTAL RESULTS

In the experiments, the QCIF version of the MISS AMERICA sequence is used for the test sequence. To compare the performance of our video coder with the current technology for low bit rate video coding (i.e., below 64 kbits/sec), we used the software simulation model of the new H.263 standard obtained from <ftp://bonde.anta.no/pub/tmn> [14].

The subband decomposition is a uniform 2×2 exact reconstruction analysis/synthesis system. We chose the recursive filter banks for their computational efficiency. The block size for searching the motion vector candidates in our algorithm is 4×4 . A search region of ± 4 pixels in both spatial directions is chosen. The threshold values T_1 and T_2 are set to be 1.2 and 3.0 respectively. The current running average Lagrangian J_a is computed based on the previous four Lagrangian J_λ values in order to make the algorithm more adaptive. All the motion vectors throughout the experiments are at whole pixel accuracy. Motion estimation is performed only for the luminance component and the estimated motion vector field was subsequently used for the motion compensation of the chrominance signals. The target bit rate is set to be approximately 16 kbits/sec.

Fig. 5 (a) and (b) show the bit rate usage and the PSNR coding performance of our coder and the H.263 standard for 50 frames of the luminance component of the color test sequence MISS AMERICA. We fixed the PSNR and compared the corresponding bit rates required by both coders.

(a) MOTION FIELD FOR FULL-SEARCH BMA
Motion Vectors of Frame 2 of Flower Garden



(b) MOTION FIELD FOR OUR ALGORITHM
Motion Vectors of Frame 2 of Flower Garden

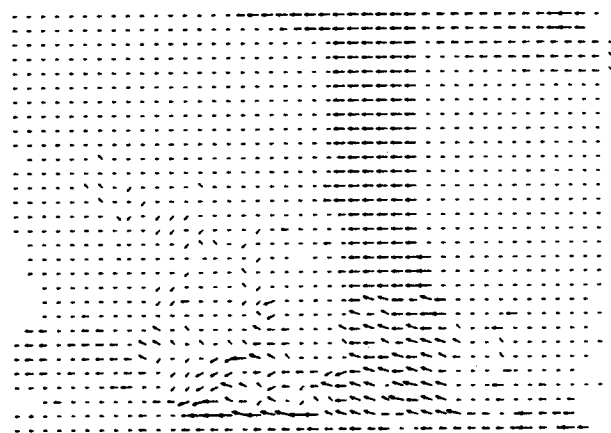


Figure 4: Motion vector field obtained by (a) FS-BMA and (b) our algorithm on the low-pass subband of FLOWER GARDEN (SIF format) Frame No. 2. We used a ± 4 pixel search region with block size of 4×4 .

While the average PSNR is approximately 39.4 dB for both coders, the average bit rate for our coder is only 13.245 kbits/sec as opposed to 16.843 kbits/sec for the H.263 standard. To achieve the same PSNR performance, our coder requires only 78% of the overall bit rate of the H.263 video coder.

Finally, to illustrate the computational reduction in motion estimation, we show the comparison in terms of number of matches required for our algorithm and the full-search BMA. For the FS-BMA with the same ± 4 search region in a subband frame, the number of matches required is $22 \times 18 \times 81 = 32076$, and $4 \times 4 = 16$ MAD calculations for each corresponding match. Our algorithm requires at most $22 \times 18 \times (9 + 9 - 1) = 6732$ matches with a table look-up of codebook size 9 with no MAD calculations. For example, in Frame 41 of the MISS AMERICA sequence, 305 out

of $22 \times 18 = 396$ matches are found in the first stage of the first layer pass. Furthermore, no match is found to exhaustively search all stages in both first and second layer passes. The algorithm uses a total of 1202 matches in comparison with 32076 matches that would have been required by the FS-BMA, and our algorithm requires no MAD calculations.

In conclusion, our algorithm outperforms the current H.263 standard by more efficient utilization of bit rates given an image quality. In terms of search complexity in motion estimation, our algorithm is able find a better motion vector field in a rate-distortion sense and requires a fraction of the computation in comparison to the full-search BMA.

5. REFERENCES

- [1] H. Musmann, "Advances in picture coding," *Proc. of the IEEE*, vol. 73, pp. 523-548, Apr. 1985.
- [2] Q. Wang and R. J. Clarke, "Motion estimation and compensation for image sequence coding," *Signal Processing: Image Communication*, no. 4, pp. 161-174, 1992.
- [3] K. I. T. Koga, A. Hirano, Y. Iijima, and T. Ishiguro., "Motion-compensated interframe coding for video conferencing," in *Proc. NTC 81*, (New Orleans), pp. G5.3.1-G5.3.5, Dec. 1981.
- [4] "Video codec for audiovisual services at $p \times 64$ kbits/s; Recommendation H.261." The International Telegraph and Telephone Consultant Committee, 1990.
- [5] Motion Picture Experts Group, ISO-IEC JTC1/SC29/WG11/602, "Generic Coding of Moving Pictures and Associated Audio," *Recommendation H.262 ISO/IEC 13818-2*, Committee Draft, Seoul, Korea, Nov. 5, 1993.
- [6] D. Anastassiou, "Current status of the MPEG-4 standardization effort," in *SPIE Proc. Visual Communications and Image Processing*, vol. 2308, pp. 16-24, 1994.
- [7] D. Le Gall, "MPEG: a video compression standard for multimedia applications," *Communications of the ACM*, vol. 34, pp. 46-58, Apr. 1991.
- [8] B. Girod, "Rate-constrained motion estimation," in *SPIE Proc. Visual Communications and Image Processing*, vol. 2308, pp. 1026-1034, 1994.
- [9] D. T. Hoang, P. M. Long, and J. S. Vitter, "Explicit bit minimization for motion-compensated video coding," in *IEEE Data Compression Conference*, (Snowbird, UT, USA), pp. 175-184, Mar. 1994.
- [10] G. J. Sullivan., "Multi-hypothesis motion compression for low bit-rate video coding," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, vol. V, pp. 437-440, 1993.
- [11] H. Gharavi, "Subband coding algorithms for video applications: Videophone to HDTV-conferencing," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 1, pp. 174-183, June 1991.
- [12] W. Chung, F. Kossentini, and M. Smith, "A new approach to scalable video coding," in *IEEE Data Compression Conference*, (Snowbird, UT, USA), Mar. 1995.
- [13] F. Kossentini, W. Chung, and M. Smith, "Image coding using high-order conditional entropy-constrained residual VQ," in *International Conference on Image Processing*, (Austin, Texas), Nov. 1994.
- [14] Telenor Research, "TMN (H.263) encoder / decoder, version 1.4a, <ftp://bonde.nta.no/pub/tmn>," *TMN (H.263) codec*, may 1995.

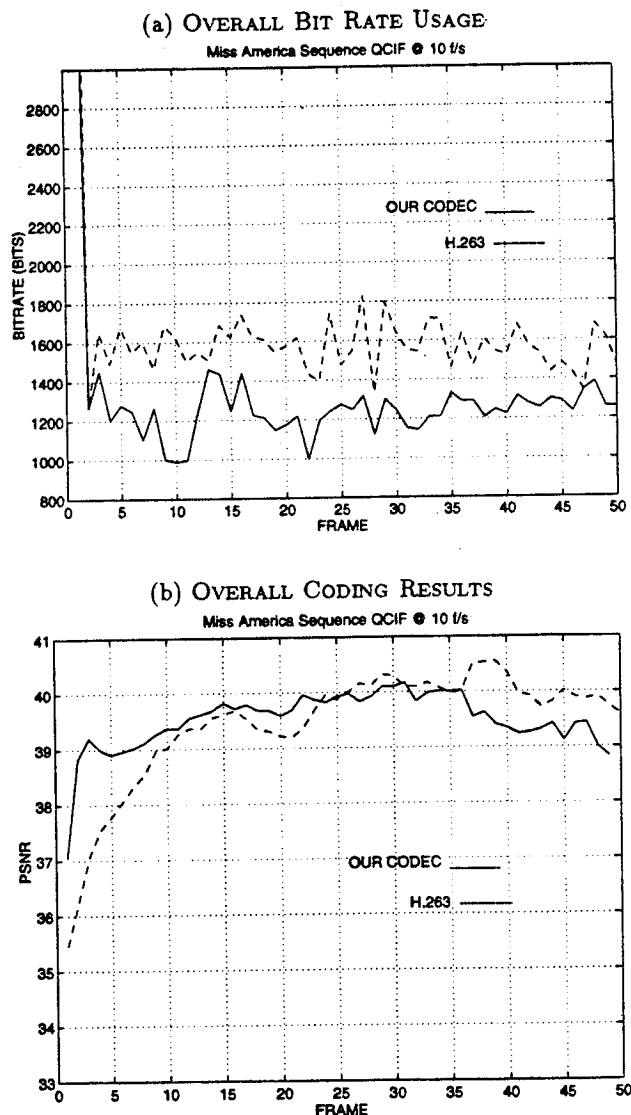


Figure 5: The comparison of overall performance — (a) bit rate usage, and (b) PSNR quality, — of our video coder with H.263 standard for MISS AMERICA QCIF sequence at 10 frames/sec. Not shown in (a) are the values 5465 bits and 7381 bits used by intra-frame (Frame 1) of our coder and H.263 standard respectively. Only the PSNR of Y luminance frames are shown in (b).